Korean Semantic Role Labeling Using CRFs

Tae-Ho Park and Jeong-Won Cha*

**Changwon National Univ., Changwon-si, Korea*
e-mail : {taehope, jcha}@changwon.ac.kr

Summary

A semantic role labeling is to select appropriate semantic relations between predicate and their arguments. There were many researches to resolve the appropriate semantic roles using syntactic information. However, there are the same semantic relations in spite of different syntactic relations and vice versa. So we have much difficulty to use syntactic information to resolve the ambiguities of semantic relations. In this paper, we suggest new features to select the appropriate semantic relations between predicates and arguments in Korean. We test the features that suggested at the previous studies, then choose the best ones for Korean. The proposed systems shows the 76.36% (F1) in the Korean Propbank and new building corpus.

Keywords: *Semantic analysis, Korean Semantic Role labeling, Conditional Random Fields*

1. Introduction

A semantic role labeling decides the semantic relation between a predicate and their arguments. A syntactic analysis finds the grammatical relations that are a subject or object etc. between words. In contrast, a semantic role labeling focus on the semantic role of arguments in sentence. The arguments have close relations between predicates and are similar to 'subject' or 'object' in syntactic analysis. There were lots of researches about decision of semantic roles using syntactic information[1-5]. However, It is difficult to determine semantic roles even if using syntactic information. Examples are as follows:

(a) 해커는 서버를 공격했다. / hae-keo-neun seo-beo-reul gong-gyeok-haet-da.

(A hacker attacked a server.)

(b) 서버는 해커에게 공격받았다. / seo-beo-neun hae-keo-e-ge gong-gyeok-ba-dat-da.

(A server was attacked by a hacker.)

'공격하다(kong-kyeok-ha-da)' can be used as '공격하다'(active voice) or '공격받다'(passive voice). The subject of (a) is '해커' and the subject of (b) is '서버'. However, the Agent(해커) and Patient(서버) of above two sentences are the same. For that reason, there are some works that use additional features like case frames on top of syntactic information[6-7]. The primary contribution of this paper is as follows: 1) testing features for domain adaptation, 2) seeking to additional features to improve SRL performance.

2. Feature analysis and selection

2.1. Baseline system

First, we measure the baseline system performance using following features:

- Complex label for syntactic information
- Parent nodes information of predicates
- Morphemes of predicate
- POS
- Combination of word, pos, dependency
- Distance between current word and their head predicates

2.1. The additional features

In this paper, we present an experimental study on solving the SRL problem in Korean using enhanced syntactic and semantic features.

- Named entities
- Phrase tags, functional tags
- Word embedding
- Word cluster using word embedding
- Grammatical information of predicates
- Semantic classes

3. Experimental result and discussion

We convert the semantic role labels defined by Sejong project to the Korean Propbank style notation. Table 1 lists the performance of the baseline system. We use 35,000 sentences originated from the Ulsan university to learn the model. We performed 5-fold cross validation experiments, given the datasets of labeled data shown in Table 2. Results show that named entity features performs better than the baseline system at ARG0, ARGM-LOC, ARGM-TMP. Especially, the performance of ARGM-TMP improved sharply because the classification between simple number expressions and tenses. We used complex labels composed of the phrase tags and functional tags in the baseline system. However, we only used the functional tags because the semantic roles are closely related to the functional tags. These features can be thought of key features to improve the performance, but we got the objection result. We used word vector using CBOW and SKIP GRAM methods. Then, we used word cluster features from the word vectors using K-means algorithm. The number of clusters of CBOW and SKIP GRAM is 100, 200 respectively. Performance rises consistently across all the cluster features; especially, when the number of cluster is 200, the performance is the best for all CBOW and SKIP GRAM method. The performance using SKIP GRAM features is better than the that of CBOW features. We used the verb derivative suffixes that can capture voice properties. We also used the semantic category to capture the semantic similarity. The semantic category can give the similarity of nouns that do not have named entity information.

Table 1. Baseline Performance

	Precision	Recall	F1
Baseline	72.53 %	67.48 %	69.99 %

4. conclusion

In this paper, we showed that some features, which are named entities, word cluster using word embedding, verb derivative suffixes and semantic category, are helpful to improve the performance in SRL. We showed in our experiments on 5-fold cross validation datasets that the proposed features are able to achieve significant improvement in performance. In future work, we will try to add new case frame features to further improve performance by capturing the property of argument omission.

Table 2. 5 Cross validation performance

	Precision	Recall	F1
Test 1	77.75	75.24	76.47
Test 2	77.60	75.26	76.41
Test 3	77.87	74.91	76.36
Test 4	77.30	74.54	75.90
Test 5	78.23	75.14	76.65
AVG	77.77	75.02	76.36

References

1. Daniel Gildea and Daniel Jurafsky. "Automatic labeling of semantic roles", Association for Computational Linguistics, Computational Linguistics, 28(3):245–288. 2002.
2. Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James Martin, and Dan Jurafsky. "Shallow semantic parsing using support vector machines. Technical", Report TR-CSLR-2003-1, Center for Spoken Language Research, Boulder, Colorado. 2003.
3. Kadri Hacioglu. "Semantic role labeling using dependency trees", In Proceedings of COLING, Geneva, Switzerland. 2004.
4. Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James Martin, and Daniel Jurafsky. "Semantic role labeling by tagging syntactic chunks", In Proceedings of CoNLL-2004, Shared Task – Semantic Role Labeling. 2004.
5. Richard Johansson and Pierre Nugues, "Dependency-based semantic role labeling of PropBank", EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing Pages 69-78, 2008.
6. Byoung-Soo Kim, Yong-Hun Lee and Jong-Hyeok Lee, "Unsupervised Semantic Role Labeling for Korean Adverbial Case", Journal of KISS : Software and Applications – 2007.6 34(2), 2007.2, 112-122, 2007.
7. Hyun-Ki Jung and Yu-Seop Kim, "Semantic Role Labeling of Korean Adverbial Arguments by using the Expanded Case Frame Dictionary", Journal of Korean Institute of Information Technology 9(10), 2011.10, 167-176, 2011.